REMARKS

The application has been reviewed in light of the Office Action dated 9 July 2004. Claims 1, 3-11, 17, 19-30 and 37-41 are presented for examination, of which Claims 1, 17, and 41 are in independent form. Claims 2, 12-16, 18, 31-36, 42 and 87-92 have been cancelled without prejudice or disclaimer of subject matter. Claims 1, 6, 17, 22, and 37-41 have been amended to define more clearly what Applicants regard as their invention, and Claim 27 has been amended to correct a typographical error. Favorable reconsideration is requested. The canceled claims will not be further addressed herein.

Claim 1 was objected to because of the reason given at page 2 of the Office Action. Applicants have amended Claim 1, among other things, to remove the single bracket. Applicants submit that the objection has been obviated, and respectfully requests its withdrawal.

Claims 37-40 were rejected under 35 U.S.C. § 102(e) as being anticipated by U.S. Patent No. 6,593,956 (*Potts et al.*); Claims 1, 3, 4, 6, 17, 19, 20, 22, 23, and 37-41 were rejected under 35 U.S.C. § 103(a) as being unpatentable over *Potts et al.* in view of U.S. Patent No. 5,995,936 (*Brais et al.*); and Claims 5, 7-11, and 24-30 were rejected under Section 103(a) as being unpatentable over *Brais et al.* as applied to Claims 1 and 17, and further in view of U.S. Patent No. 5,500,671 (*Andersson et al.*)

As shown above, Applicants have amended independent Claims 1, 17, and 41 in terms that more clearly define what they regard as their invention. Applicants submit that these amended independent claims, together with the remaining claims dependent thereon, are patentably distinct from the cited prior art for at least the following reasons.

The aspect of the present invention set forth in Claim 1 is an apparatus for processing image data and sound data. The apparatus includes an image processor, an

input unit, a data store, a sound processor, a speaker identifier, a voice recognition parameter selector, and a voice recognition processor. The image processor processes image data recorded by at least one camera showing the movements of a plurality of people to track the respective position of each person in three dimensions. The input unit inputs voice recognition parameters for each person. The data store stores data assigning a unique identity to each person tracked by the image processor and stores respective voice recognition parameters for each person. The sound processor processes sound data to determine the direction of arrival of the sound. The speaker identifier determines the unique identity of the person who is speaking by comparing the positions of the people determined by the image processor and the direction of arrival of the sound determined by the sound processor to identify a person at a position in the direction from which the sound arrives. The voice recognition parameter selector selects the voice recognition parameters from the data store of the person identified by the speaker identifier to be the person who is speaking, and the voice recognition processor processes the received sound data to generate text data therefrom using the voice recognition parameters selected by the voice recognition parameter selector.

By virtue of the these features, the apparatus of Claim 1 is able to generate data for improved archiving and subsequent retrieval of the image data. More particularly, a person who is speaking in an image can be uniquely identified, speech recognition parameters unique to that person can be recovered from memory, and the recovered speech recognition parameters can be used for voice recognition processing to generate text data to record the words spoken by that person.

To achieve this, data is stored assigning a unique identity to each person appearing in the image data, together with respective voice recognition parameters for each

person. The image data is processed to track the respective position of each person in three-dimensions. When sound data is detected representing speech, the direction of arrival of the sound is determined. The positions of the people determined by processing the image data are then compared to the determined direction of arrival of the sound to uniquely identify a person who is speaking. This is performed by determining which of the people has a 3D position corresponding to a position from which the sound has been generated (see, for example page 34 line 1 - page 35 line 20 in the present application).[1] Using the stored data uniquely identifying each person in the image data and linking each person to their respective voice recognition parameters, the voice recognition parameters of the identified speaker are retrieved from memory. The retrieved parameters are then used in voice recognition processing to convert the received sound data to text data for storage along with the image data.

Applicants submit that the applied art, alone or in combination, is not seen to disclose or suggest the invention as defined by independent Claim 1.

Potts et al. relates to determining the location of a speaker and to pan, tilt and zoom a camera to frame a better camera image of the speaker (see, for example, column 6, lines 27-33). To achieve this, the *Potts et al.* system uses an audio-based locator 70 to process audio signals received by a microphone array to determine the location and distance of a speaker relative to the microphone array (column 7, lines 35-38, column 18, lines 24-28). Audio-based locator 70 then generates a series of camera positioning directives to pan, tilt and zoom camera 14 to appropriately frame the speaker within images recorded by the camera (column 7, lines 38-40, column 18, lines 61-63, column 19, lines

---

[1]/It is to be understood, of course, that the claim scope is not limited by the details of the described embodiments, which are referred to only to facilitate explanation.

53-56). Camera 14 is then moved in accordance with the positioning directives generated by audio-based locator 70 (column 7, lines 43-46, column 19 lines 53-56). After the camera is moved, video-based locator 60 captures a frame of video image data from camera 40 and detects the two-dimensional location of any faces in the video image (column 7, lines 45-51, column 19, lines 59-62). Audio speaker validation module 116 determines whether a detected face is located in an area in the video frame where the image of the speaker's face is expected, the center of the video frame. If a face is not located at the expected position, it is assumed that the face located closest to the expected position is the face of the speaker (column 20, lines 29-36, column 21, lines 19-29, and Fig. 16). Video offset/error measurement module 104 within the video-based locator 60 determines the offset of the detected face from the center of the camera image (column 7, lines 59-66, column 12, lines 13-29, column 20, lines 36-38, column 21, lines 29-34). Camera control module 80 then uses the results to correct for the offset (column 7, lines 52-54, column 8, lines 55-59, column 19, lines 62-65, column 20, lines 19-27, column 21, lines 34-41). Data from the video-based locator 60 may also be used to correct errors in the distance of the speaker calculated by audio-based locator 70 (column 21, lines 42-45). To do this, the single face determined by audio speaker validation module 116 to be the face of the speaker is processed to compare its size with a predetermined size. Any differences in size are used to correct the camera zoom directives (column 21, line 57 to column 22, line 5).

Brais et al. relates to automated report generation in dependence upon speech to text processing, spoken commands, relative time information, and captured video images or clips.

As previously stated, Applicants submit that the features of the Claim 1 are not disclosed or suggested by the applied art, and in particular *Potts et al.* and *Brais et al.*

Applicants submit that nothing has been found in *Potts et al.* that would teach or suggest an image processor which processes image data recorded by at least one camera showing the movements of a plurality of people to track the respective position of each person in three-dimensions. In the *Potts et al.* system, a two-dimensional position of faces in a frame of video data is detected, a single face is selected as the face of the speaker, and a range of the single face is calculated. However, there is no tracking of the three-dimensional positions of a plurality of people performed in the *Potts et al.* system.

Further, Applicants submit that nothing has been found in *Potts et al.* that would teach or suggest a data store storing data assigning a unique identity to each person tracked by the image processor and to store respective voice recognition parameters for each person. Paragraph 17 of the Office Action acknowledges that *Potts et al.* does not teach a storage unit for storing voice recognition parameters for each of a plurality of people. In addition, however, *Potts et al.* also fails to disclose or suggest storing data assigning a unique identity to each person tracked by the image processor. More particularly, the *Potts et al.* system detects the face of a speaker but never determines the identity of that face. The consequence of this is that the *Potts et al.* system can never be used to identify a speaker and select speech recognition parameters for that person because the *Potts et al.* system never determines to whom a detected face belongs.

Furthermore, nothing has been found in *Potts et al.* that would teach or suggest a speaker identifier determining the unique identity of the person who is speaking by comparing the positions of the people determined by the image processor and the direction of arrival of the sound determined by the sound processor to identify a person at a position in the direction of the sound. In contrast, the *Potts et al.* system works in a very different way. In the *Potts et al.* sytem, the position of the speaker is first determined based

on the direction of arrival of the sound. This position is then used to move the camera so that the speaker is expected to be at the center of an image recorded by the camera. Only after the camera has been moved, is the image data processed to detect the two-dimensional position of faces therein. The face at the position closest to the center of the image is selected to be the face of the speaker.

Still further, nothing has been found in *Potts et al.* that would teach or suggest a voice recognition parameter selector selecting the voice recognition parameters from the data store of the person identified by the speaker identifier to be the person who is speaking, and a voice recognition processor operable to process the received sound data to generate text data therefrom using the voice recognition parameters selected by the voice recognition parameter selector. Paragraph 16 of the Office Action acknowledges that *Potts et al.* does not teach a voice recognition processor, while paragraph 17 acknowledges that neither *Potts et al.* nor *Brais et al.* teaches a selector for selecting voice recognition parameters. As explained above, the *Potts et al.* system is simply not suitable to provide an input to a voice recognition parameter selector because the *Potts et al.* system does not assign a unique identity to the detected face of a speaker (and merely detects the face without determining to whom it belongs).

For at least the above reasons, Applicants submit that a combination of *Potts et al.* and *Brais et al.*, assuming such combination would even be permissible, would fail to teach or suggest the features of Claim 1.

Accordingly, Applicants submit that Claim 1 is patentable over the cited art, and respectfully request withdrawal of the rejection under 35 U.S.C. § 103(a).

Independent Claim 17 is a method claim corresponding to apparatus Claim 1, and is believed to be patentable over *Potts et al.* and *Brais et al.* for at least the same

reasons as discussed above in connection with Claim 1. Additionally, independent Claim 41 includes features similar to those discussed above in connection with Claim 1. Accordingly, Claim 41 is believed to be patentable for reasons substantially similar as those discussed above in connection with Claim 1.

A review of the other art of record has failed to reveal anything which, in Applicants' opinion, would remedy the deficiencies of the art discussed above, as references against the independent claims herein. Those claims are therefore believed patentable over the art of record.

The other claims in this application are each dependent from one or another of the independent claims discussed above and are therefore believed patentable for the same reasons. Since each dependent claim is also deemed to define an additional aspect of the invention, however, the individual reconsideration of the patentability of each on its own merits is respectfully requested.

In view of the foregoing amendments and remarks, Applicants respectfully request favorable reconsideration and early passage to issue of the present application.

Applicants' undersigned attorney may be reached in our New York office by telephone at (212) 218-2100. All correspondence should continue to be directed to our below listed address.

Respectfully submitted,

Ronald A. Clayton
Attorney for Applicants
Registration No.: 26,718

FITZPATRICK, CELLA, HARPER & SCINTO
30 Rockefeller Plaza
New York, New York 10112-3801
Facsimile: (212) 218-2200

NY_MAIN 456680v1